Effects of ℓ_2-regularization and hyperparameters on loss landscape

Haruka Eshima, Makoto Yamada

Machine Learning and Data Science Unit, OIST Graduate University, Okinawa, Japan

https://oist.mlds.jp haruka.eshima@oist.jp

Introduction

- **Model.** Two-layer neural network (NN).
- **Background.** Different training dynamical regimes (lazy vs non-lazy) are observed both empirically and theoretically. [1] [2]
- **Motivation.** Explain the training dynamics differences from a perspective on the loss landscape of finite-width NNs.
- **Contributions.**
  - ▷ Geometrical difference between the set of globally optimal parameters for squared loss with or without ℓ_2-regularization.
  - ▷ Relationship between the location of globally optimal parameters and the location of initial parameters.
- **Conclusion.** Hyperparameter space classification boundary for training dynamics at infinite width found by [1] is explained from a perspective on loss landscape analysis.

Background

- **Lazy Regime vs Non-lazy Regime.**
  For a study on training dynamics for two-layer fully connected neural networks

$$f_\theta(x) \propto \sum_{j=1}^{m} \sigma(\langle u_j, x\rangle)\omega_j,$$

  infinite-width setting $m \to \infty$, continuous-time limit, loss is differentiable w.r.t. f_θ such as squared loss

$$\frac{d\theta}{dt} \propto -\nabla_\theta \hat{\mathcal{R}}_n(\theta(t)), \quad \hat{\mathcal{R}}_n(\theta(t)) = \frac{1}{2n}\sum_{i=1}^{n}(f_\theta(x_i) - y_i)^2$$

  are often used to describe parameter dynamics. [1] [2]
  Previous research [1] drew a phase diagram of hyperparameters for a two-layer ReLU NN at the infinite-width limit for a characterization of training dynamical regimes, namely, $\sup_{t\in[0,\infty)} RD(u(t)) \to 0$, $\mathcal{O}(1)$, or ∞ as $m \to \infty$, where

$$RD(u(t)) = \frac{\|u(t) - u(0)\|_2}{\|u(0)\|_2}.$$

  - ▷ What happens if we add ℓ_2-regularization?
  - ▷ How about in a finite-width setting?
- **Loss Landscape Analysis.**
  The questions above are considered in loss landscape analysis. [3]
  - ▷ Can we relate loss landscape and training dynamics?

Problem Setup

- **Model.**
  Two-layer ReLU NN for one-dimensional data.

$$f_\theta(x) = \frac{1}{\alpha}\sum_{j=1}^{m}(xu_j)_+\omega_j.$$

  $\alpha = m^{a_1}$ is the scaling factor
  The parameters are initialized independently by $u_i^0,\ \omega_i^0 \sim N(0, \tau^2)$ and $\tau = m^{-b_1}$.
  We assume that m is sufficiently large.
- **Loss Functions.**

$$l(\theta) = \frac{1}{2}\left\|\frac{1}{\alpha}\sum_{j=1}^{m}(Xu_j)_+\omega_j - Y\right\|_2^2 \quad \text{and}$$

$$L(\theta) = \frac{1}{2}\left\|\frac{1}{\alpha}\sum_{j=1}^{m}(Xu_j)_+\omega_j - Y\right\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(u_j^2 + \omega_j^2).$$

  $\beta = m^{-c_1}$ is the weight decay coefficient and $c_1 > a_1$.

Methods

- **Exact derivation of global optima.**

$$\varphi^*(m) = \left\{(u_j, \omega_j)_{j=1}^m \,\middle|\, \sum_{j:\,u_j\geq 0} u_j\omega_j = \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}, \sum_{j:\,u_j\leq 0} u_j\omega_j = \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}\right\}.$$

$$\Theta^*(m) = \left\{(u_i, \omega_i)_{i=1}^m \,\middle|\, \sum_{i:\,u_i\geq 0} u_i^2 = |\gamma_P^*|, \sum_{i:\,u_i\leq 0} u_i^2 = |\gamma_N^*|,\right.$$

$$\left.\omega_i = \begin{cases} \mathrm{sign}(\gamma_P^*)\, u_i & (u_i > 0), \\ \mathrm{sign}(\gamma_N^*)\, u_i & (u_i < 0), \\ 0 & (u_i = 0) \end{cases}\right\}$$

- **Geometrical Properties.** (Under mild assumptions.)
  - ▷ Without ℓ_2-regularization,

$$\forall a \geq \frac{a_1}{2},\ \exists \varphi_a^*(m) \subset \varphi^*(m)\ s.t.\ \forall \phi_a^* \in \varphi_a^*(m),$$
$$\|\phi_a^*\|_2 = \Theta(m^a) \text{ and } \dim(\varphi_a^*(m)) = 2m - 2.$$

  - ▷ With ℓ_2-regularization,

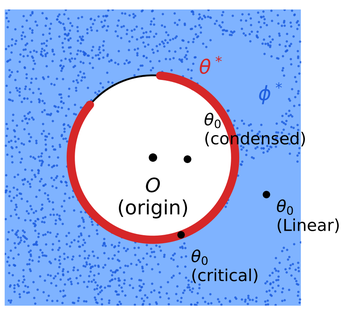
$$\|\theta^*\|_2 = \Theta\left(m^{\frac{a_1}{2}}\right) \text{ and } \dim(\Theta^*(m)) = m - 2.$$

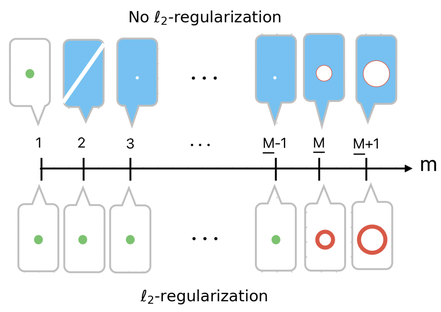
- **Concentration of the initial parameter location.**

$$\mathbb{P}\left(\frac{\sqrt{2}}{2}m^{\frac{1-2b_1}{2}} \leq \|\theta_0\| \leq \frac{3\sqrt{2}}{2}m^{\frac{1-2b_1}{2}}\right) \geq 1 - 2\exp\left(-\frac{cm}{2\kappa^2}\right).$$

Results

- **Effects of ℓ_2-regularization.**
  - ▷ Adding the ℓ_2-regularization term reduces the dimension of the set of optimal parameters by m and imposes a bound on the set.
- **Effects of hyperparameters.**
  - ▷ Hyperparameter space classification boundary based on the relationship between the locations of initialized parameters and global optima matches with the boundary based on training dynamics behaviors at infinite width found by [1].





Discussions

- **Next Questions**.
  - ▷ Does lazy regime exist for loss with ℓ_2-regularization?
  - ▷ Can loss landscape analysis explain the change in training dynamics over training time?
  - ▷ Can we extend our analysis to general d-dimensional input data?

[1] Luo, T., Xu, Z.-Q. J., Ma, Z., and Zhang, Y. Phase diagram for two-layer relu neural networks at infinite-width limit. *JMLR*, 2021.
[2] Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *NeurIPS*, 2019.
[3] Kim, S., Mishkin, A., and Pilanci, M. Exploring the loss landscape of regularized neural networks via convex duality. *ICLR*, 2025.